DOCUMENT RESUME

ED 160 636 TM 007 720

AUTHOR Divgi, D. R.
TITLE A New Index for the Accuracy of a Criterion-Referenced Test.
PUB DATE Mar 78
NOTE 12p.; Paper presented at the National Council on Educational Measurement (Toronto, Ontario, Canada, March 27-30, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Complexity Level; *Criterion Referenced Tests; *Cutting Scores; *Mathematical Models; *Standard Error of Measurement; Test Items; *Test Reliability; True Scores

ABSTRACT

One aim of criterion-referenced testing is to classify an examinee without reference to a norm group; therefore, any statements about the dependability of such classification ought to be group-independent also. A population-independent index is proposed in terms of the probability of incorrect classification near the cutoff true score. The compound binomial model leads to the conclusion that a criterion-referenced test is more reliable if the item difficulties are unequal. The proposed index also has the following properties: (1) A test with a higher value of r will result in a smaller loss due to misclassification; (2) It is easily generalized to a case of more than two classes; and (3) It has no theoretical upper limit, but increases with test length. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D. R. Divgi

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

ED160636

A NEW INDEX FOR THE ACCURACY OF A CRITERION-REFERENCED TEST

D. R. Divgi

Syracuse University

Abstract

One aim of criterion-referenced testing is to classify an examinee without reference to a norm group. Therefore any statements about the dependability of such classification ought to be group-independent also. A population-independent index is proposed in terms of the probability of incorrect classification near the cutoff true score. The compound binomial model leads to the conclusion that a criterion-referenced test is more reliable if the item difficulties are unequal.

Paper presented at the annual meeting of the National Council on Educational Measurement, Toronto, March 27-30, 1978.

TM007 720

Printed in U.S.A.

A NEW INDEX FOR THE ACCURACY OF A

CRITERION-REFERENCED TEST

The purpose of a criterion-referenced test is to classify each examinee into one of m mastery states. The variability of scores on such a test can be quite small, leading to a small value of the classical reliability coefficient even when the classifications based on the test are quite accurate. (Popham and Husek, 1969) Therefore, for the case of m = 2, Livingston (1972) modified the classical definition to obtain a new index given by

$$\frac{\mathrm{var}(\tau) + (\mu - C)^2}{\mathrm{var}(X) + (\mu - C)^2} = \frac{\rho + (\mu - C)^2/\mathrm{var}(X)}{1 + (\mu - C)^2/\mathrm{var}(X)}$$

where τ and X are true and observed scores respectively, μ is the mean score, C is the criterion score, and ρ the classical coefficient of reliability. For any given value of ρ, Livingston's index increases with the distance between the mean and criterion scores, and therefore it can have an appreciable value even when $\rho = 0$.

Another argument against using the classical coefficient is that it is based on differences among all individuals, whether they are placed in the same class or not. However the basic aim of criterion-referenced testing, in contrast with norm-referenced testing, is to avoid comparisons among individuals beyond the unavoidable ones involved in classifying them. Therefore differences among scores of examinees placed in the same class are of no interest. (Hambleton and Novick, 1973) From this viewpoint it becomes natural to inquire about the consistency of decisions based on two different administrations of the same test (or of parallel forms). Swaminathan, Hambleton, and Algina (1974) advocated using Cohen's (1960)

kappa index as the measure of consistency. Huynh (1976) has made a detailed study of the kappa index estimated from a single administration using the beta-binomial model. Subkoviak (1976) has used the proportion of individuals with the same classification on both occasions as the measure of consistency. He, too, uses the binomial error model but does not make any assumptions about the distribution of true scores. Instead he uses the regression estimate of the true score for each individual.

All these indices depend on the distribution of ability in the group tested. This goes against the basic goal of making a statement about each individual without any reference to a norm group. Such a statement is not complete without some specification of the error of measurement or of the reliability of the test. If true group-independence is to be achieved, one's measure of reliability should be group-independent also.

Another problem is that there is at present no criterion for choosing among different indices, particularly between kappa and the coefficient of agreement, both of which arise from the decision-theoretic approach. Kappa as a function of the criterion score is largest when the cutoff is near the group mean (Huynh, 1976, 1977), but the coefficient of agreement is smallest in such a case (subkoviak, 1976, Fig. 1), as is Livingston's index. Shall we say a test is most reliable or least reliable when the cutoff score is close to the group mean?[1]

---

[1]Huynh (1977) proposes that one should avoid the issue of dependence on the criterion score by using the maximum value of kappa (called kappamax or KM). One can similarly use the corresponding minimum values of the other indices. Then Livingston's index reduces to the classical reliability ρ; kappamax is closely approximated by $(2/\pi)\sin^{-1}\hat{\rho}$. (Huynh 1977, Table 2) Thus one is back full circle to using the classical coefficient of reliability.

A NEW INDEX

Obviously no such conflict will arise if one uses an index which does not depend on the population. For a norm-referenced test, if one can specify the error variance as a function of true score, one has a detailed description of measurement error which is independent of the population. In case of a criterion-referenced test the quantity analogous to the standard error of measurement is the probability of incorrect classification. A complete group-independent description consists of specifying this probability as a function of true score.

First one needs a definition of correct classification. For simplicity let us consider the case of only two classes. Let ζ be the individual's proportion-correct true score. He will be classified as a Master (Nonmaster) if his true score is larger (smaller) than a criterion value ζ_c. (Hambleton and Novick, 1973) Determination of ζ_c on theoretical or empirical grounds is very difficult. We shall assume that the cutoff score C is given, and an examinee with C or more items correct is classified as a Master. In classical theory true score is defined as the mean of observed scores over a large number of administrations of the same test under identical conditions. Analogously we shall define correct classification as that which the individual obtains more frequently, i.e., the one which is more probable for that individual. Therefore the criterion true score ζ_c is that value of ζ at which the two classifications are equally likely. Then the correct classification is Master if $\zeta > \zeta_c$ and Nonmaster if $\zeta < \zeta_c$. For any given value of ζ, one can calculate the probability of incorrect classification using a mathematical model.

Provided the model is appropriate, one can obtain a curve showing the probability of error P_e as a function of ζ. This curve will have a cusp at ζ_o with the ordinate equal to .5. P_e will decrease monotonically with distance from ζ_c on either side. Such a curve provides a complete description of the error in decisions based on the test. However a single number is much more convenient. The natural candidate, which does not require any arbitrary choices, is the rate at which P_e decreases from its maximum value of .5 at ζ_c.

$$P(x \geq C \mid \zeta) = 1 - P(x < C \mid \zeta)$$

Therefore our index of accuracy is

$$r = \lim_{\epsilon \to 0} \left[.5 - P_e(\zeta_c + \epsilon)\right] / \epsilon$$

$$= \lim_{\epsilon \to 0} \left[.5 - P_e(\zeta_c - \epsilon)\right] / \epsilon$$

$$= \frac{d}{d\zeta} P(x \geq C \mid \zeta) \qquad \text{at } \zeta = \zeta_c .$$

The value of this derivative will depend on the values of the parameters of the model. A larger value of the derivative implies smaller values of P_e in the neighborhood of ζ_c. It seems plausible that any change that reduces the probability of error for ζ near ζ_o will do so at all values of ζ. (This has been verified for the compound binomial model which is described below.) Therefore a higher value of r implies a smaller probability of incorrect classification at any true score, and therefore a smaller expected loss for any given distribution of true scores in the population. No such direct connection with expected loss function has been demonstrated for any of the indices mentioned above.

This index is easily generalized to the case of m classes. There are (m-1) cutoff scores. For each of these one can determine the true proportion score where the two adjacent classes are equally likely. Then one computes the value of r at this criterion true score. One may average these (m-1)

values or use each of them separately to judge the ability of the test to distinguish between the corresponding adjacent classes.

The index r is not a ratio of two quantities, and it has no theoretical upper limit. It can be infinite for an infinitely long test. Therefore it does not seem appropriate to call it an index of reliability. One may define

$$r' = 1 - 1 / (1 + r^2).$$

This lies between 0 and 1 but has no simple interpretation.

AN ILLUSTRATION

Let us now illustrate the above procedure using Lord's (1965) compound binomial model. In this model the probability of score x on a test of n items is given by

$$P(x|\zeta) = p_n(x) + k\,\zeta(1-\zeta)\,C_2(x)$$

where

$$p_n(x) = \binom{n}{x} \zeta^x (1-\zeta)^{n-x} \qquad \text{if } 0 \leq x \leq n,$$

$$= 0 \qquad \text{otherwise;}$$

$$C_2(x) = \sum_{u=0}^{2} (-1)^{u+1} \binom{2}{u} p_{n-2}(x-u).$$

k is a parameter of the model which can be estimated by setting the observed KR-20 equal to its theoretical value. This yields

$$k = \frac{n^2 (n-1) s_p^2}{2\left[\bar{x}(n-\bar{x}) - s_x^2 - n s_p^2\right]}$$

where s_p^2 is the variance of item difficulties, $\bar{x}$ is the mean score, and s_x^2 the variance of observed scores. The value of k is expected to remain stable from one group to another provided ζ measures the same trait in all the groups.

Thus k is a property of the test itself, irrespective of the ability distribution in the group used to estimate its value.

Given the test length n, cutoff score C, and the parameter k, one can use the Newton-Raphson procedure to determine the value ζ_c of ζ where the probability of obtaining a score $x \geq C$ equals .5. This turns out to be slightly smaller than (2C-1)/2n. Then one computes the derivative $dP(x \geq C|\zeta)/d\zeta$ at this point to obtain the value of r. Table 1 contains the results for test lengths n = 10, 20, and 30 with C = .8n, and various values of k. The value of r increases with test length as $\sqrt{n}$, just as the reciprocal of the standard error of measurement of x/n does. As k is almost directly proportional to s_p^2, the monotonic increase in r with k means that a test is more accurate for purposes of classification if difficulties of items are different rather than equal. (This conclusion cannot be carried to large values of k because then the model yields negative "probabilities" of appreciable magnitudes. This limitation is not serious since the items in a criterion-referenced test have a narrowly defined domain of content, and hence cannot differ much in difficulty.) The relationship between accuracy and k is easy to understand in view of the fact that the error variance is given by

$$\text{var}(x|\zeta) = (n - 2k)\zeta(1 - \zeta)$$

(Lord, 1965, eqn. 34) Increase in k reduces the error of measurement and hence the probability of misclassification. The results in Table 2 show that this happens not only near the criterion score but at all values of ζ. The dependence of r on the cutoff score is shown in Table 3. The index is largest when the cutoff score has the largest or smallest possible value.

To summarize, the proposed index of accuracy for a criterion-referenced test has the following properties.

(a) It is independent of the group to which the test is administered.

(b) It requires a mathematical model. However this model need not be complicated. A one-parameter model is enough to lead to a definite prescription for improving accuracy.

(c) Assume that, for given test length and cutoff score, whatever reduces error at values of ζ near ζ_c has a similar effect at all values of ζ. Then, for any given distribution of true scores, a test with a higher value of r will result in a smaller expected loss due to misclassification.

(d) It is easily generalized to a case of more than two classes.

(e) It has no theoretical upper limit. It increases with test length as $\sqrt{n}$. Thus it is analogous to the reciprocal of the standard error of measurement of the proportion correct score, rather than to the coefficient of reliability.

REFERENCES

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Hambleton, R. K. & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Huynh, H. The kappamax consistency index for decisions in domain-referenced testing. Paper presented at the annual meeting of the American Educational Research Association, New York, 1977.

Livingston, S. A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26.

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, 239-270.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Subkoviak, M. J. Estimating reliability from a single administration of a mastery test. Journal of Educational Measurement, 1976, 13, 265-276.

Swaminathan, H., Hambleton, R. K., & Algina, J. J. Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-267.

Table 1

Dependence of the index of accuracy on test length and on parameter k

10k/n	n = 10 C = 8	n = 20 C = 16	n = 30 C = 24
0.0	2.962	4.298	5.315
0.2	3.021	4.384	5.421
0.4	3.081	4.470	5.527
0.6	3.140	4.556	5.633
0.8	3.200	4.642	5.740
1.0	3.260	4.728	5.846
1.2	3.320	4.815	5.953
1.4	3.381	4.902	6.059
1.6	3.441	4.988	6.166
1.8	3.501	5.075	6.272
2.0	3.562	5.162	6.379

Table 2

Probabilities of incorrect classification at different values of k and of true proportion-correct score ξ for a test with n = 10, C = 8

	Values of ξ					
k	0.4	0.5	0.6	0.7	0.8	0.9
0.0	.012	.055	.167	.383	.322	.070
1.0	.004	.035	.139	.362	.316	.049
2.0	.000	.016	.110	.341	.309	.028

Table 3

Variation of index r with values of cutoff score and of parameter k for a test with 10 items

Cutoff	k = 0.0	k = 1.0	k = 2.0
1	5.364	5.937	6.574
2	3.540	3.900	4.274
3	2.962	3.259	3.562
4	2.703	2.975	3.246
5	2.598	2.857	3.117
6	2.598	2.857	3.117
7	2.704	2.975	3.246
8	2.962	3.260	3.562
9	3.539	3.899	4.272
10	5.366	5.939	6.574